Differential Enriched Scan 2 (DEScan2): a fast pipeline for broad peak analysis

Righelli D(1,2)[†] , Koberstein J(3), Gomes B(4), Zhang N(5), Angelini C(1), Peixoto L(4), Risso Davide(6).

(1) Institute for Applied Mathematics "M. Picone", National Research Council, Napoli.
(2) Department of Business Sciences, Management and Innovation Systems , University of Salerno, Salerno.
(3) Oregon Health and Science University, Portland, Oregon
(4) Elon S. Floyd College of Medicine, Washington State University, Spokane, Washington
(5) Statistics Department, Wharton University of Pennsylvania, Philadelphia, Pennsylvania
(6) Division of Biostatistics and Epidemiology, Weill Cornell Medicine, New York, New York

[†] Email: d.righelli@na.iac.cnr.it

Motivation
Next Generation Sequencing (NGS) techniques revolutionized biology enabling to examine biological processes by different points of view producing a vast amount of data. In this context, the most widely investigated aspects are the transcriptional level with RNA-Seq, the epigenetic state with ChIP-Seq and BS-seq and the chromatin accessibility with Sono-Seq/Atac-Seq. Nowadays, the analysis of RNA-seq, ChIP-seq and BS-Seq can be considered well established, whereas the analysis of Sono-Seq/Atac-Seq is still challenging. Despite the lack of robust computational methods for their analysis, recent studies have demonstrated the relevance of Sono-Seq/Atac-Seq to unveil the significant role of open state regions of chromatin linked to diseases like autism [1], among many others. To fill the gap in existing methods, we present DEScan2 a novel bioconductor package [2] for the analysis of Sono-Seq/Atac-Seq data.
Methods
The method consists of three main steps: 1) a peak caller, 2) a peak filtering and 3) a method to efficiently compute a count matrix of the filtered peaks. The peak caller in step 1) is a standard moving window scan that compares the counts within a sliding window to the counts in a larger region outside the window, using a simple Poisson likelihood (no overdispersion estimation) and providing a final z-score for each detected peak. However, the package can work with any external peak caller returning results in terms of bed files, indeed the package provides additional functionalities to load bed files of peaks and handle them as GenomicRanges structures [3]. The filtering step 2 is aimed to determine if a peak is a "true peak" on the basis of its replicability in other samples. Basing on this idea, we developed the filtering step to filter out those peaks not present in at least a given number of samples. In the light of this, the user can decide the minimum number of samples where each peak

has to be detected. Moreover, a further threshold can be used over the peak score. Finally, the third step produces a count matrix where each column is a sample and each row a filtered peak computed in the filtering step. The value of the matrix cell is the number of reads for the peak in the sample. Furthermore, our package provides several functionalities for GenomicRanges data structure handling. One over the others gives the possibility to split a GenomicRanges over the chromosomes to speed-up the computations parallelizing them over the chromosomes.

Results

We illustrate the performance of our pipeline using two published datasets for chromatin accessibility. The first example is a Sono-Seq dataset [1] describing the chromatin accessibility of the mouse hippocampus following fear conditioning (8 samples, 4 per each condition), in order to reproduce their analysis on which our pipeline is designed.

The second example [4] describes adult mouse dentate granule neurons in vivo before and after synchronous neuronal activationusing Atac-Seq and RNA-Seq technologies, of which we selected the first 8 samples, 4 per each condition. Since in this case also RNA-Seq are available we will illustrate how our Atac-Seq pipeline can be placed in a more generic and complex process for NGS data integration. To this purpose we considered two integration methods mixOmics [5] and MoFa [6] which use as main data structure for their integration a count matrix, making our pipeline extremely suited for this purpose. Finally, we will briefly dicuss few ideas on how to extend our method to the analysis of not yet published single-cell dataset on Atac-Seq data.

References

1. KOBERSTEIN, J.N., et al. Learning-dependent chromatin remodeling highlights noncoding regulatory regions linked to autism. Sci. Signal., 2018, 11.513: eaan6500.

2. https://doi.org/doi:10.18129/B9.bioc.DEScan2

3. LAWRENCE, M, et al. Software for Computing and Annotating Genomic Ranges.. PLoS Computational Biology, 2013.

4. SU, Y., et al. Neuronal activity modifies the chromatin accessibility landscape in the adult brain. Nature neuroscience, 2017, 20.3: 476.

5. DEJEAN, S., et al. mixOmics: Omics data integration project. R package, 2013.

6. ARGELAGUET, R., et al. Multi-Omics factor analysis disentangles heterogeneity in blood cancer. bioRxiv 217554;